**Health Weather: Real Time Analysis of Flu Outbreaks Using Social Media**

*Sai Kishore Bandaru, Jagadish Rao, Ebenezer Anand Arapally, Priyanka*

**Abstract:**

“Health is Wealth” is the popular slogan we hear all the time. Now a days certain flus are spreading rapidly and also with less awareness. But with the increase in the usage of social media, the news about such flus and its symptoms were being tweeted or posted on the social media pages. Using the current social media we can track the flus going in any place. Keeping these points in view, we are planning develop a plan which will help in tracking the flus in a place and also find the remedies for the issue.

**Introduction:**

The project typically deals with real time analysis of tweets which are collected which are precise to health issues. But here we will have two parts in the project. One part deals with the collection of tweets regarding flus. Another part deals with analyzing of static trained data which contains symptoms and respective treatment done at different hospitals. Now we will combine both this parts and provide a system to find the outbreaks and find out respective treatment or measures to be taken for cure. In this paper we shall discuss about the feature extraction that is being done as part of the project.

The below figure depicts the architecture of the system:

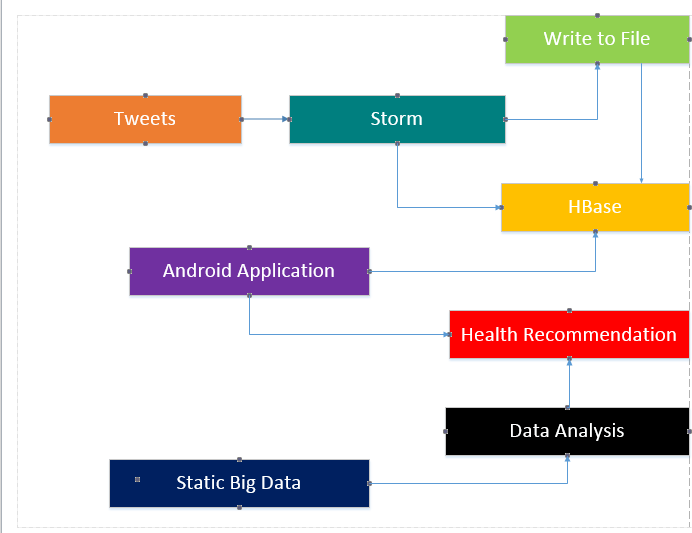


Figure 1System Architecture

In the first part we will be collecting data from the Twitter using STORM. Now using storm we will be collecting tweets and storing in a file. Now using the bolts in storm architecture we will be storing the tweets from file to HBase. Now we will be retrieving the real time data from Hbase either by web service or by HttpClient methodologies. Now to make recommendations to the user based on flu outbreaks, we use the static big data we have collected and analyze those data using the R java. Using R we shall use either naïve Bayes classification or K-means cluster to make recommendations for the users. Our application is a native mobile application developed for android users.

So here we collected the data from storm and use this real time data with trained data to analyze the tweets. Here we used Naïve Bayes, D- Tree, SVM methodologies to find which analysis algorithm to be used for accuracy. We have designed the ROC curves and taken the accuracy numbers to find the optimum classification algorithm. The Android mobile application typically prints the confusion matrix of the selected algorithm.

**Materials and Methods:**

We have used twitter as source for real time data. We used twitter4j APIs to retrieve the data from twitter. Now each bolt in the STORM will be retrieving this information and will be storing in a temporary buffer and this data will be made available for each bolt. Where one bolt will classify the real time data based on trained data using the Naive Bayes classifier. Now for the classification process we have used two features one is negative and positive outbreak.

Cleaning Data: The tweets contains many unnecessary syntactic features. The cleaning is done such that the data confirms WEKA’s data format.

The following is removed from the tweet message

* -quotes
* -@ symbol
* -# symbol
* -URL
* -‘RT’ word
* -smiles

Additionally duplicates are also removed to reduce unnecessary computational overload.

**Results:**

We have observed significant results for decision tree. But to evaluate the results the trained data could have been much better. The data which we have taken and classified has been limited. So future evaluation will be based on more features which includes a minimum of five features and the classifier used will be decided based on accuracy and on the ROC curves. . The following are the related ROC curves and confusion matrix for different classification algorithms:

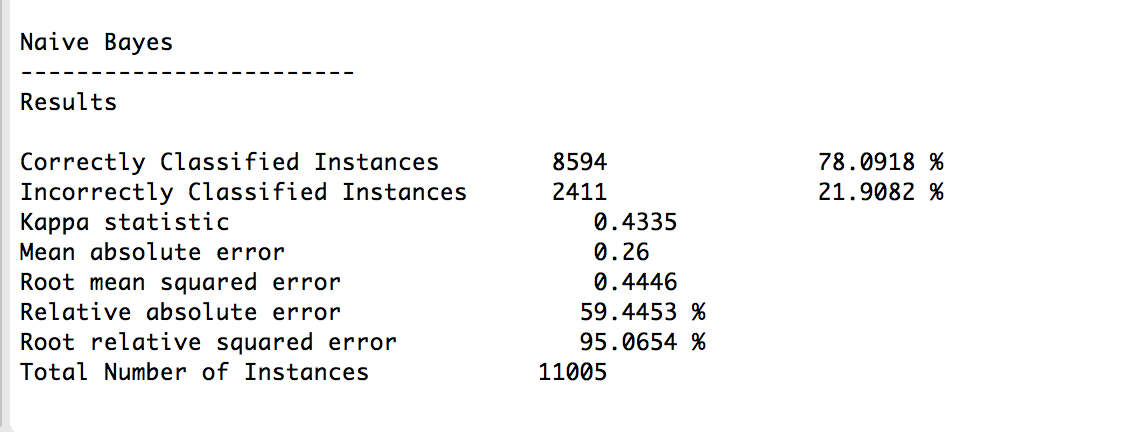
****

Figure 2 Confusion Matrix for Naive Bayes classification

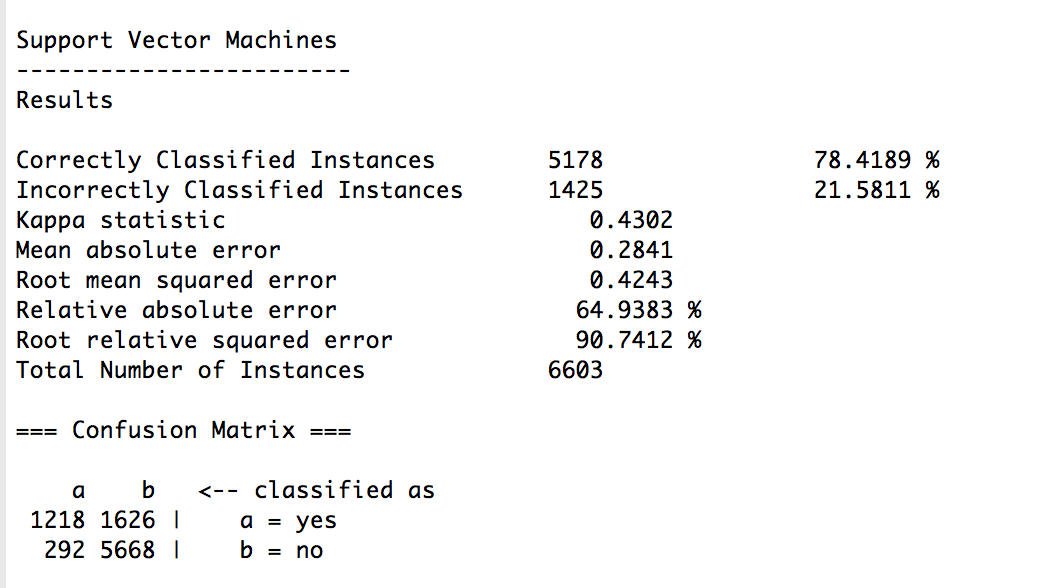


Figure 3SVM classification

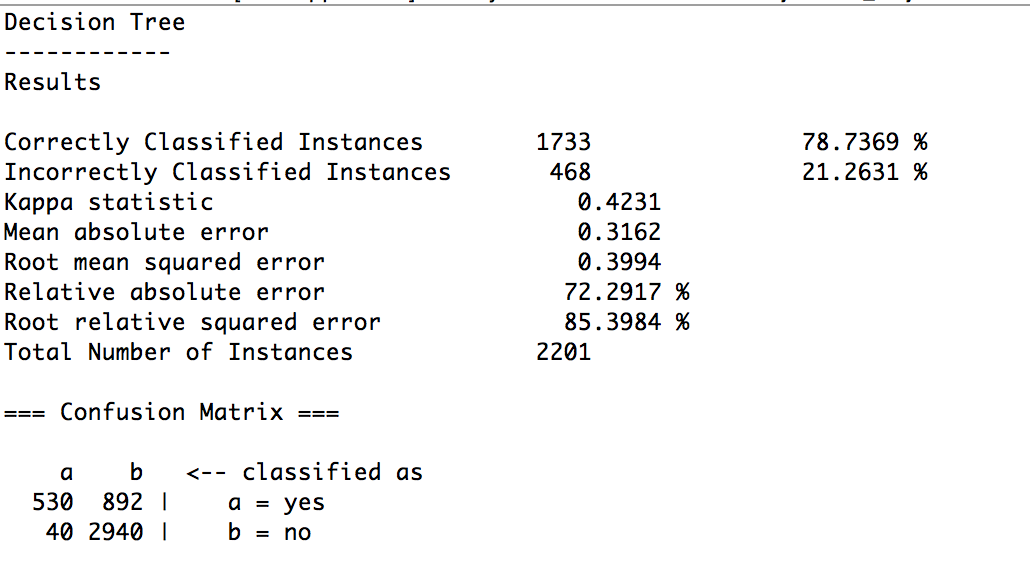


Figure 4Decision Tree confusion matrix

The another form of methodology we took for depicting the accuracy are the ROC curves where we can trace the true positive and true negative values based on the data and also on the classification. The following figure shows the ROC curves for the three classification algorithms:

|  |  |  |
| --- | --- | --- |
| C:\Users\SaiKishore\Desktop\Screen Shot 2014-11-03 at 11.10.09 PM.png | C:\Users\SaiKishore\Desktop\Screen Shot 2014-11-03 at 11.13.08 PM.png | C:\Users\SaiKishore\Desktop\Screen Shot 2014-11-03 at 11.15.23 PM.png |

ROC curves for Naïve Bayes, Decision Tree, SVM classification algorithms

We have developed a Native mobile application for Android users. The mobile interface has a button which on click displays the confusion matrix for a model. The following are the snapshots of the control flow:

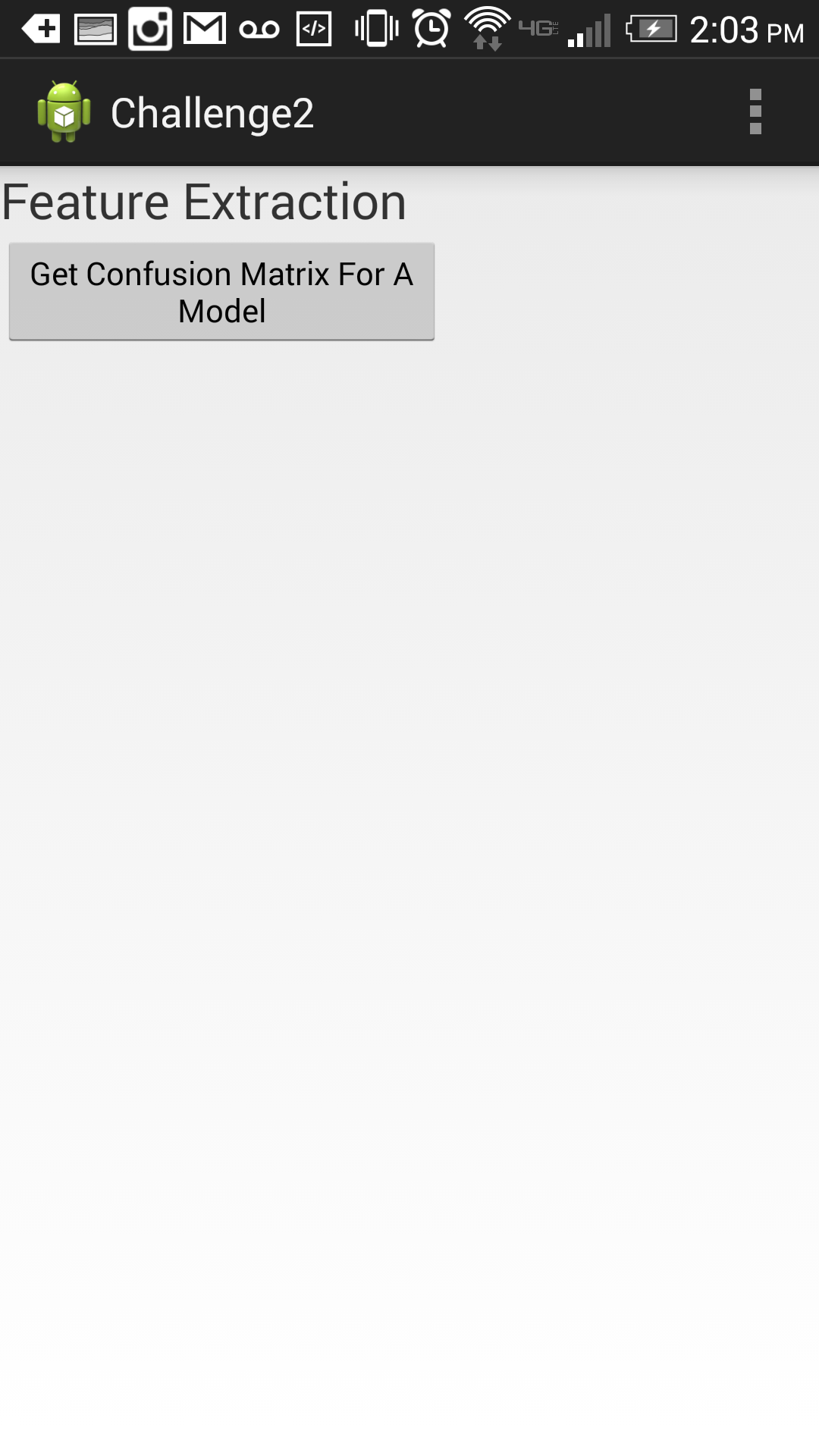


Figure 5Initial WebPage

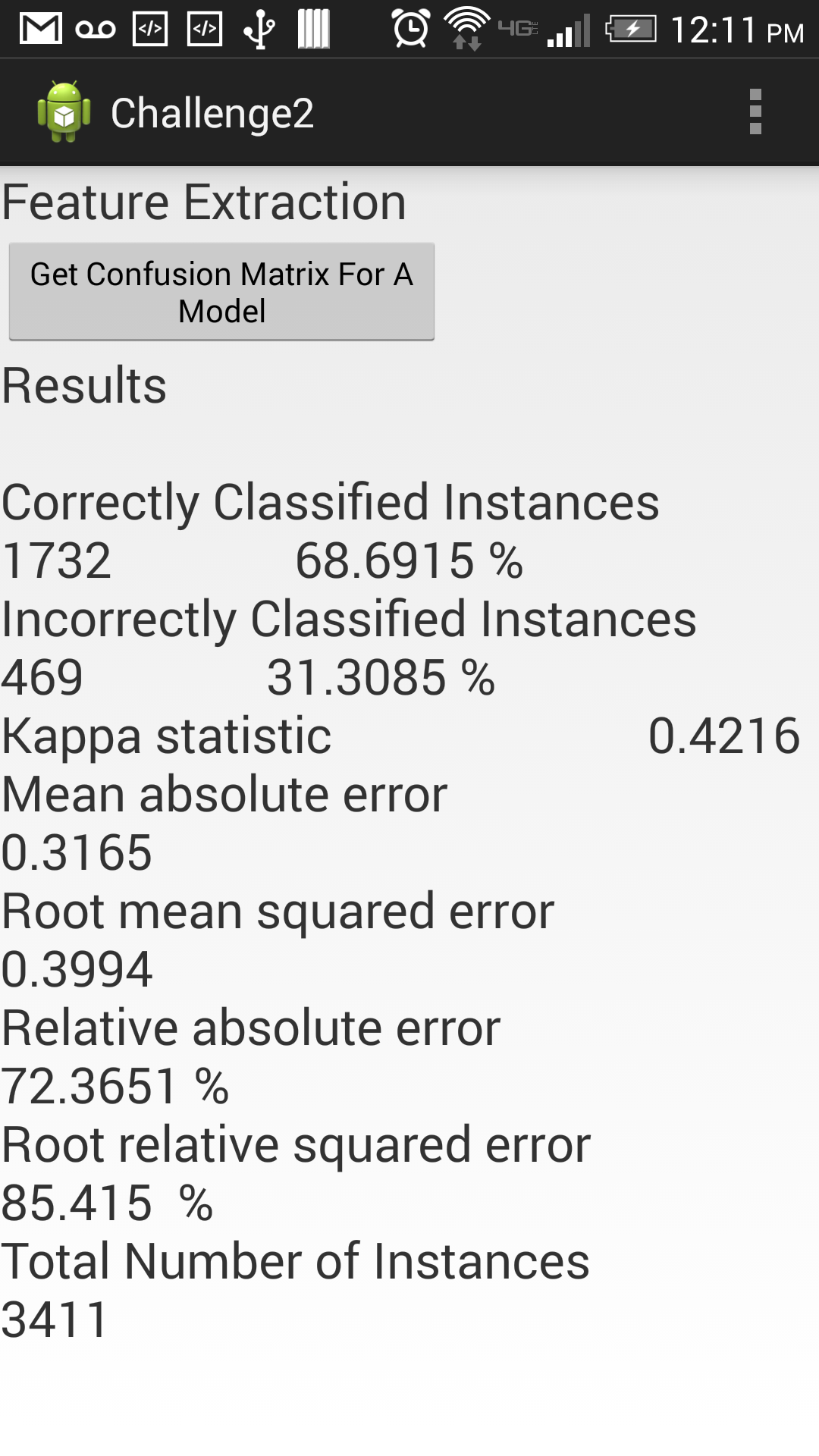


Figure 6Confusion matrix

**Discussions:**

As we observed in the previous sections that decision tree algorithm has given more accurate results when compared to other classification algorithms. However when we tried doing the evaluation we spent lot of time for the data pruning and formatting part while making the evaluation for the decision tree classification. Since we couldn’t observe significant difference in the data accuracy for Naive Bayes and decision tree we decided to use the Naïve Bayes as our classification algorithm. But since the evaluation paved a way and since we have are having a well compiled executed set of code we are planning to change the data and also the features for the flu outbreak evaluation rather being precise to positive and negative outbreaks.

**Conclusions:**

Based on our evaluation process for feature extraction we observed that the data which has been collected should be trained properly and also neatly quoted. The classification algorithm Naïve Bayes proved to be more feasible and relatively accurate when compared to other algorithms.

**Acknowledgments:**

This project is being done as partial fulfillment of the academic course Real Time and Analytics at University of Missouri Kansas City, under the esteemed guidance of:

* Instructor: Dr.Yugyung Lee
* TA: Feichen Shen

**Literature Cited:**

* L. Barbosa and J. Feng. Robust sentiment detection on Twitter from biased and noisy data. In Proceedings of the 23rd International Conference on Computational Linguistics: Posters. Association for Computational Linguistics, 2010.
* M. Bautin, L. Vijayarenu, and S. Skiena. International sentiment analysis for news and blogs. In Proceedings of the International Conference on Weblogs and Social Media (ICWSM), 2008.
* V.L. Corpora. Empirical Methods in Natural Language Processing.
* N. Godbole, M. Srinivasaiah, and S. Skiena. Large-scale sentiment analysis for news and blogs. In Proceedings of the International Conference on Weblogs and Social Media (ICWSM). Cite seer, 2007.